TAAI 2024

A Study of LLM Generated Pseudo-Data for Improving Small-Scale Models in Human Values Estimation

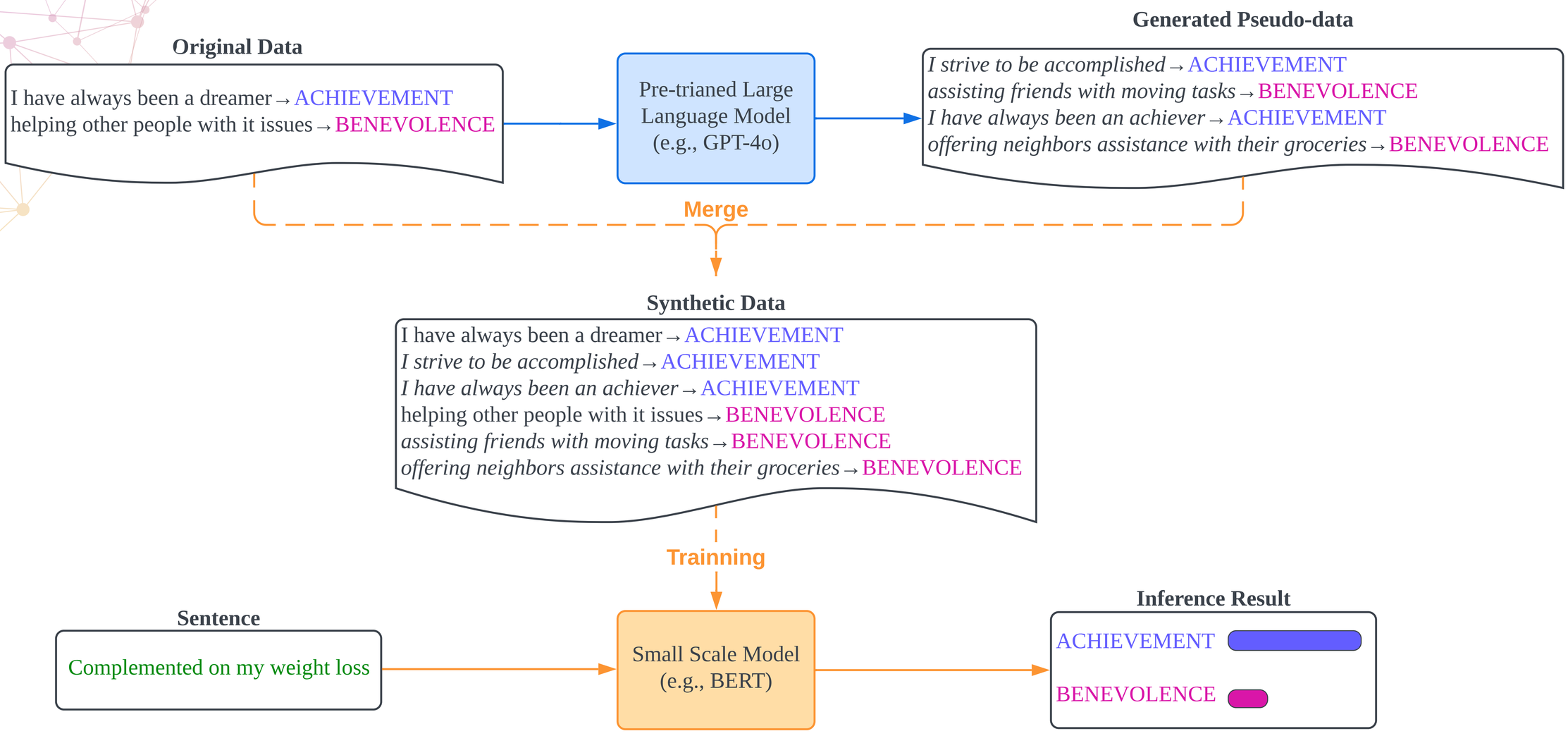
Yihong Han, Rintaro Tomitaka, Yoko Nishihara, Megumi Yasuo, Junjie Shan
Ritsumeikan University

Main Contributions

- Evaluate effectiveness of LLM-generated pseudo-data in augmenting human values datasets
- Compare performance between small-scale models and direct LLM approaches
- Analyze impact of varying pseudo-data proportions on model performance

Proposed Method

- ValueNet dataset as original data
- GPT-4o as pre-trained LLM
- BERT base uncased as small scale model
- Generate pseudo data from human value definition and sample data.
- Build synthetic dataset from 1x to 4x of the original size, category balanced



ValueNet data number

Human Values	Original
ACHIEVEMENT	192
BENEVOLENCE	888
CONFORMITY	91
HEDONISM	819
POWER	438
SECURITY	637
SELF-DIRECTION	108
STIMULATION	305
TRADITION	98
UNIVERSALISM	294
Total	3870

Experimental Results

Experiment case	Accuracy
LLM zero-shot	0.25
LLM few-shot	0.27
Original dataset	0.4
size = 1x(balanced)	0.45
size = 2x	0.53
size = 3x	0.565
size = 4x	**0.57**

Key Findings

- Synthetic dataset brought 17% accuracy improvement (0.4→0.57)
- Proposed method outperformed LLM-only approach (0.57 vs 0.25 & 0.27)
- A balanced dataset brought 5% accuracy improvement (0.4→0.45)
- Accuracy improvement became minor after dataset size over 3

Conclusion

- The proposed method achieved an accuracy improvement in human values estimation
- Small-scale models trained with synthetic data outperformed LLM-only approaches
- Early stages of data augmentation showed the most substantial performance improvements
- Accuracy improvement in the late stage is minor.
- Balanced dataset creation through pseudo-data generation helped address the data scarcity